Querying Attributed DL-Lite Ontologies Using Provenance Semirings

Camille Bourgaux, Ana Ozaki

▶ **To cite this version:**

Camille Bourgaux, Ana Ozaki. Querying Attributed DL-Lite Ontologies Using Provenance Semirings. Thirty-Third AAAI Conference on Artificial Intelligence (AAAI-19), Jan 2019, Honolulu, United States. hal-02109645

HAL Id: hal-02109645
https://inria.hal.science/hal-02109645

Submitted on 25 Apr 2019

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Querying Attributed DL-Lite Ontologies Using Provenance Semirings

Camille Bourgaux
Télécom ParisTech & DI ENS, CNRS, ENS,
PSL University & Inria, France

Ana Ozaki
KRDB Research Centre,
Free University of Bozen-Bolzano, Italy

Abstract

Attributed description logic is a recently proposed formalism, targeted for graph-based representation formats, which enriches description logic concepts and roles with finite sets of attribute-value pairs, called annotations. One of the most important uses of annotations is to record provenance information. In this work, we first investigate the complexity of satisfiability and query answering for attributed DL-Lite$_\mathcal{R}$ ontologies. We then propose a new semantics, based on provenance semirings, for integrating provenance information with query answering. Finally, we establish complexity results for satisfiability and query answering under this semantics.

Introduction

Description logic (DL) (Baader et al. 2007) ontologies allow to express complex relationships between concepts and roles, but they are ill-equipped to represent and reason about multiple and heterogeneous types of meta-knowledge, such as the temporal validity of a fact, or its source. For instance, the YAGO ontology (Hoffart et al. 2013) attaches provenance metadata to its facts (e.g., source and confidence of the extraction) as well as temporal and geospatial information. Many practical applications therefore use *knowledge graphs*, which consist, like DL assertions, of directed labelled graphs but that also allow, unlike DLs, to add annotations to vertices and edges. Property Graph, the data model used in many graph databases (Rodriguez and Neubauer 2010), and Wikidata, the knowledge graph used by Wikipedia (Vrandečić and Krötzsch 2014), are prominent examples of such labelled graphs. To bridge the gap between DL and knowledge graphs, *attributed description logics* (Krötzsch et al. 2017; Krötzsch et al. 2018) have been recently introduced. They enrich DL concepts and roles with finite sets of attribute-value pairs, called *annotations*, and allow to express constraints on these annotations in the ontology inclusions. For example, the attributed DL assertion $\mathsf{spouse}(\mathsf{taylor}, \mathsf{burton})@[\mathsf{start} : 1975, \mathsf{end} : 1976]$ states that Liz Taylor was married to Richard Burton from 1975 to 1976, and the following role inclusion expresses that spouse is a symmetric relation, where the inverse statement has the same start and end dates:

$$\mathsf{spouse}@X \sqsubseteq \mathsf{spouse}^-@[\mathsf{start} : X.\mathsf{start}, \mathsf{end} : X.\mathsf{end}].$$

While the work by Krötzsch et al. studied the complexity of the satisfiability problem for several attributed DL languages, our focus in this paper is on *query answering* in attributed DL. The problem of querying DL ontologies using database-style queries (in particular, conjunctive queries) is an important research topic for which tractable DL languages have been tailored (Bienvenu and Ortiz 2015). We consider here the DL-Lite$_\mathcal{R}$ dialect of the *DL-Lite family* (Calvanese et al. 2007), which underlies the OWL 2 QL profile (Motik et al. 2009), and investigate attributed DL-Lite$_\mathcal{R}$.

One of the main motivations of attributed DLs is to integrate annotations carrying provenance information, which are very frequent in knowledge graphs[1]. Recording and tracking provenance information is an important topic in database theory, where *provenance semirings* (Green, Karvounarakis, and Tannen 2007) were introduced as an abstract tool to relate the result of a query with information about the original sources of the data and the ways in which the query was obtained. Such information comes in the form of a *provenance polynomial*. It has been useful for many applications, such as query answer explanation or querying of probabilistic databases (Senellart 2017; Cheney, Chiticariu, and Tan 2009; Suciu et al. 2011). Bienvenu, Deutch, and Suchanek (2012) argued that provenance would be useful for Web data, e.g., to establish the authorship or determine the trust in a given piece of data, or to help to guarantee the privacy of information. Provenance has also been investigated for non-relational databases and Semantic Web (see Conclusion for discussion of related work). In this work, we propose a new semantics for the attributed DL annotations, based on provenance semirings, so that queries can be annotated with provenance polynomials. To the best of our knowledge, this is the first work where provenance polynomials are embedded into both the syntax and the semantics of the query.

The first section introduces attributed DL-Lite$_\mathcal{R}$, following the formalism given by Krötzsch et al. (2017; 2018). We then define attributed conjunctive queries and study the complexity of satisfiability and query answering in attributed DL-Lite$_\mathcal{R}$. We next present our new semantics for the annotations to model provenance and analyse the complexity of satisfiability and query answering with this new model, considering queries that can be annotated with provenance polynomials. In particular, we show that satisfiability and query answering

[1] E.g., in Wikidata *reference* (provenance) is one the most frequent types of annotations `https://www.wikidata.org`.

in attributed DL-Lite$_\mathcal{R}$ are PSPACE-complete problems. For the semirings-based semantics and queries annotated with provenance polynomials, we establish that although satisfiability is EXPTIME-hard in the general case, the new semantics does not increase the complexity of query answering if the ontology contains only assertions and a restricted form of inclusions, which is close to the database setting considered by Green, Karvounarakis, and Tannen (2007). We also investigate various restrictions of the general setting. Our results are for *combined complexity*, when both the query and the ontology are considered as the input. Proofs are available in our technical report (Bourgaux and Ozaki 2018).

Attributed DL-Lite

Attributed DLs are defined over the usual DL signature with countable sets of *concept names* $\mathsf{N_C}$, *role names* $\mathsf{N_R}$, and *individual names* $\mathsf{N_I}$. We consider an additional set $\mathsf{N_U}$ of *set variables* and a set $\mathsf{N_V}$ of *object variables*. Annotation sets are defined as finite binary relations, understood as sets of attribute-value pairs. Attributes and values refer to domain elements and are syntactically denoted by individual names. To describe annotation sets, we use *specifiers*. The set $\mathbf{S}$ of specifiers contains the following expressions:

- set variables $X \in \mathsf{N_U}$;
- *closed specifiers* $[a_1 : v_1, \dots, a_n : v_n]$; and
- *open specifiers* $\lfloor a_1 : v_1, \dots, a_n : v_n \rfloor$,

where $a_i \in \mathsf{N_I}$ and v_i is either an individual name in $\mathsf{N_I}$, an object variable in $\mathsf{N_V}$, or an expression of the form $X.a$, with X a set variable in $\mathsf{N_U}$ and a an individual name in $\mathsf{N_I}$. We use $X.a$ to refer to the (finite, possibly empty) set of all values of attribute a in an annotation set X. A *ground specifier* is a closed or open specifier that only contains individual names. Intuitively, closed specifiers define specific annotation sets whereas open specifiers merely provide lower bounds (Krötzsch et al. 2017).

Syntax. A DL-Lite$^\mathcal{R}_@$ *role* (resp. *concept*) *assertion* is an expression $R(a,b)@S$ (resp. $A(a)@S$), with $R \in \mathsf{N_R}$ (resp. $A \in \mathsf{N_C}$), $a, b \in \mathsf{N_I}$, and $S \in \mathbf{S}$ a ground closed specifier. DL-Lite$^\mathcal{R}_@$ *role* and *concept inclusions* are of the form $X : S$ $(P \sqsubseteq Q)$ and $X : S$ $(B \sqsubseteq C)$ respectively, where $X \in \mathsf{N_U}$, $S \in \mathbf{S}$ is a closed or open specifier, and P, Q and B, C are respectively role and concept expressions defined by the following syntax, where $A \in \mathsf{N_C}$, $R \in \mathsf{N_R}$ and $S \in \mathbf{S}$:

$$P ::= R@S \mid R^-@S, \qquad Q ::= P \mid \neg P,$$
$$B ::= A@S \mid \exists P, \qquad C ::= B \mid \neg B.$$

We further require that all variables are *safe*. For set variables, this means that if $Y \in \mathsf{N_U}$ occurs on the right side of an inclusion (or in a specifier S such that the prefix of the inclusion is $X : S$ and X occurs on the right side), then the specifier of the left side expression is Y. For object variables, if they occur on the right side of an inclusion then they must also occur on the left side or in S such that $X : S$ and X occurs on the left. Note that if object variables occur in S with $X : S$ in the prefix and X on the right side, then X is the specifier on the left by the safety definition. If the prefix of an inclusion is $X : S$ and X does not occur in the role/concept expressions of the inclusion, we may ommit $X : S$.

A DL-Lite$^\mathcal{R}_@$ *ontology* is a set of DL-Lite$^\mathcal{R}_@$ assertions, role and concept inclusions. Also, we say that a DL-Lite$^\mathcal{R}_@$ ontology is *ground* if it does not contain variables. To simplify notation, we omit the specifier $\lfloor\rfloor$ (meaning "any annotation set") in role or concept expressions. In this sense, any DL-Lite$_\mathcal{R}$ axiom is also a DL-Lite$^\mathcal{R}_@$ axiom. Moreover, we omit prefixes of the form $X : \lfloor\ \rfloor$, which state that there is no restriction on X. The *size* of an ontology $\mathcal{O}$ (or a query, defined later), which we may denote with $|\mathcal{O}|$, is the length of the string that represents it.

Example 1. *Our running example's ontology $\mathcal{O}_{\mathsf{ex}}$ expresses that those who are married (role* spouse*) to someone are married (concept* Married*), annotated with the same sources from which the information has been extracted (attribute* src*):*

$$\exists \mathsf{spouse}@X \sqsubseteq \mathsf{Married}@\lfloor \mathsf{src} : X.\mathsf{src} \rfloor.$$

The assertion states that Zsa Zsa Gabor was married to Jack Ryan and it is annotated with the sources of this information:

$$\mathsf{spouse}(\mathsf{gabor}, \mathsf{ryan})@[\mathsf{src} : \mathsf{s}_1, \mathsf{src} : \mathsf{s}_2].$$

Semantics. An *interpretation* $\mathcal{I} = (\Delta^\mathcal{I}, \cdot^\mathcal{I})$ of an attributed DL consists of a non-empty domain $\Delta^\mathcal{I}$ and a function $\cdot^\mathcal{I}$. Individual names $a \in \mathsf{N_I}$ are interpreted as elements $a^\mathcal{I} \in \Delta^\mathcal{I}$. To interpret annotation sets, we use the set $\Phi^\mathcal{I} := \{\Sigma \subseteq \Delta^\mathcal{I} \times \Delta^\mathcal{I} \mid \Sigma \text{ is finite}\}$ of all finite binary relations over $\Delta^\mathcal{I}$. Each concept name $A \in \mathsf{N_C}$ is interpreted as a set $A^\mathcal{I} \subseteq \Delta^\mathcal{I} \times \Phi^\mathcal{I}$ of elements with annotations, and each role name $R \in \mathsf{N_R}$ is interpreted as a set $R^\mathcal{I} \subseteq \Delta^\mathcal{I} \times \Delta^\mathcal{I} \times \Phi^\mathcal{I}$ of pairs of elements with annotations. Each element (pair of elements) may appear with multiple different annotations.

$\mathcal{I}$ *satisfies* a concept assertion $A(a)@[a_1 : v_1, \dots, a_n : v_n]$ if $(a^\mathcal{I}, \{(a_1^\mathcal{I}, v_1^\mathcal{I}), \dots, (a_n^\mathcal{I}, v_n^\mathcal{I})\}) \in A^\mathcal{I}$. Role assertions are interpreted analogously. Expressions with free set or object variables are interpreted using variable assignments $\mathcal{Z}$ mapping object variables $x \in \mathsf{N_V}$ to elements $\mathcal{Z}(x) \in \Delta^\mathcal{I}$ and set variables $X \in \mathsf{N_U}$ to finite binary relations $\mathcal{Z}(X) \in \Phi^\mathcal{I}$. For convenience, we also extend variable assignments to individual names, setting $\mathcal{Z}(a) = a^\mathcal{I}$ for every $a \in \mathsf{N_I}$. A specifier $S \in \mathbf{S}$ is interpreted as a set $S^{\mathcal{I},\mathcal{Z}} \subseteq \Phi^\mathcal{I}$ of matching annotation sets. We set $X^{\mathcal{I},\mathcal{Z}} := \{\mathcal{Z}(X)\}$ for variables $X \in \mathsf{N_U}$. The semantics of closed specifiers is defined as:

- $[a : v]^{\mathcal{I},\mathcal{Z}} := \{\{(a^\mathcal{I}, \mathcal{Z}(v))\}\}$ where $v \in \mathsf{N_I} \cup \mathsf{N_V}$;
- $[a : X.b]^{\mathcal{I},\mathcal{Z}} := \{\{(a^\mathcal{I}, \delta) \mid (b^\mathcal{I}, \delta) \in \mathcal{Z}(X)\}\}$;
- $[a_1 : v_1, \dots, a_n : v_n]^{\mathcal{I},\mathcal{Z}} := \{\bigcup_{i=1}^n F_i \mid F_i \in [a_i : v_i]^{\mathcal{I},\mathcal{Z}}\}$.

$S^{\mathcal{I},\mathcal{Z}}$ therefore is a singleton set for set variables and closed specifiers. For open specifiers, however, we define $\lfloor a_1 : v_1, \dots, a_n : v_n \rfloor^{\mathcal{I},\mathcal{Z}}$ to be the set:

$$\{F \subseteq \Phi^\mathcal{I} \mid F \supseteq G \text{ for } \{G\} = [a_1 : v_1, \dots, a_n : v_n]^{\mathcal{I},\mathcal{Z}}\}.$$

Now given $A \in \mathsf{N_C}$, $R \in \mathsf{N_R}$, and $S \in \mathbf{S}$, we define:

$$(A@S)^{\mathcal{I},\mathcal{Z}} := \{\delta \mid (\delta, F) \in A^\mathcal{I} \text{ for some } F \in S^{\mathcal{I},\mathcal{Z}}\},$$
$$(R@S)^{\mathcal{I},\mathcal{Z}} := \{(\delta, \epsilon) \mid (\delta, \epsilon, F) \in R^\mathcal{I} \text{ for some } F \in S^{\mathcal{I},\mathcal{Z}}\}.$$

Further DL expressions are defined as usual: $(R^-@S)^{\mathcal{I},\mathcal{Z}} = \{(\gamma,\delta) \mid (\delta,\gamma) \in (R@S)^{\mathcal{I},\mathcal{Z}}\}$, $\neg P^{\mathcal{I},\mathcal{Z}} = (\Delta^{\mathcal{I}} \times \Delta^{\mathcal{I}}) \setminus P^{\mathcal{I},\mathcal{Z}}$, $\exists P^{\mathcal{I},\mathcal{Z}} = \{\delta \mid \text{there is } (\delta,\epsilon) \in P^{\mathcal{I},\mathcal{Z}}\}$, $\neg C^{\mathcal{I},\mathcal{Z}} = \Delta^{\mathcal{I}} \setminus C^{\mathcal{I},\mathcal{Z}}$. $\mathcal{I}$ *satisfies* a concept inclusion $X : S$ $(B \sqsubseteq C)$ if, for all variable assignments $\mathcal{Z}$ that satisfy $\mathcal{Z}(X) \in S^{\mathcal{I},\mathcal{Z}}$, we have $B^{\mathcal{I},\mathcal{Z}} \subseteq C^{\mathcal{I},\mathcal{Z}}$. Satisfaction of role inclusions is defined analogously. An interpretation $\mathcal{I}$ satisfies an ontology $\mathcal{O}$, or is a *model* of $\mathcal{O}$, if it satisfies all of its axioms. As usual, $\models$ denotes the induced logical entailment relation.

Example 2 (Example 1 cont'd)**.** *Consider an interpretation $\mathcal{I}$ with domain $\Delta^{\mathcal{I}} = \{\mathsf{gabor}, \mathsf{ryan}, \mathsf{src}, \mathsf{s}_1, \mathsf{s}_2\}$ and such that $\cdot^{\mathcal{I}}$ maps each individual name to itself and*

$$\mathsf{spouse}^{\mathcal{I}} = \{(\mathsf{gabor}, \mathsf{ryan}, \{(\mathsf{src}, \mathsf{s}_1), (\mathsf{src}, \mathsf{s}_2)\})\}$$

$$\mathsf{Married}^{\mathcal{I}} = \{(\mathsf{gabor}, \{(\mathsf{src}, \mathsf{s}_1), (\mathsf{src}, \mathsf{s}_2)\})\}.$$

The interpretation $\mathcal{I}$ is a model of $\mathcal{O}_{\mathsf{ex}}$.

Reasoning in DL-Lite$^{\mathcal{R}}_{@}$

In this section, we study the complexity of satisfiability and query answering over DL-Lite$^{\mathcal{R}}_{@}$ ontologies. Our first result is that the satisfiability problem, which is in NL for DL-Lite$_{\mathcal{R}}$ (Artale et al. 2009), is harder for DL-Lite$^{\mathcal{R}}_{@}$. The proof is by reduction from the word problem for polynomially space bounded deterministic Turing Machines (DTM). Annotations raise the complexity because they can encode configurations of a DTM, using expressions of the form $X.b$ to encode the synchronization of successive configurations.

Theorem 1. *In DL-Lite$^{\mathcal{R}}_{@}$, satisfiability is* PSPACE*-hard.*

To prove the PSPACE upper bound for satisfiability, we use *grounding* (Krötzsch et al. 2017), which is a classical technique that consists in eliminating variables from an ontology to transform it into an equisatisfiable ground ontology. The ground ontology can then be translated into an equisatisfiable DL-Lite$_{\mathcal{R}}$ ontology. The grounding leads to an exponential blowup of the ontology while the translation to DL-Lite$_{\mathcal{R}}$ is polynomial. Since satisfiability of DL-Lite$_{\mathcal{R}}$ ontologies is in NL (Artale et al. 2009), it follows (by (Savitch 1970)) that satisfiability of DL-Lite$^{\mathcal{R}}_{@}$ ontologies is in PSPACE.

Theorem 2. *In DL-Lite$^{\mathcal{R}}_{@}$, satisfiability is in* PSPACE.

We now turn our attention to the problem of querying DL-Lite$^{\mathcal{R}}_{@}$ ontologies. In the following we only define and deal with conjunctive queries without free variables, i.e., boolean conjunctive queries (BCQ), as the problem of finding certain answers to a query is reducible to BCQ entailment.

Definition 1 (Attributed Queries)**.** *An* attributed boolean conjunctive query *(BCQ$_@$) q is an expression of the form:*

$$\exists \boldsymbol{x}.X_1 : S_1, \ldots, X_n : S_n \; \varphi(\boldsymbol{x}) \tag{1}$$

where, for $1 \leq i \leq n$, X_i are the set variables occurring in $\varphi(\boldsymbol{x})$, $S_i \in \mathbf{S}$, and $\varphi(\boldsymbol{x})$ is a conjunction of atoms of the form $A(t)@S$ or $R(t,u)@S$, with $A \in \mathsf{N_C}$, $R \in \mathsf{N_R}$, $S \in \mathbf{S}$, and t, u individual names in $\mathsf{N_I}$ or variables in $\boldsymbol{x} \subseteq \mathsf{N_V}$.

We may write $E(\boldsymbol{t})@S$ to refer to an atom of any of the two forms ($E \in \mathsf{N_C} \cup \mathsf{N_R}$ and $\boldsymbol{t}$ is a tuple of elements from $\mathsf{N_I} \cup \boldsymbol{x}$ of the arity of E).

An interpretation $\mathcal{I} = (\Delta^{\mathcal{I}}, \cdot^{\mathcal{I}})$ satisfies a BCQ$_@$ *q, written $\mathcal{I} \models q$, if there exists a variable assignment $\mathcal{Z}$ such that:*

- *$\mathcal{Z}(X_i) \in S_i^{\mathcal{I},\mathcal{Z}}$ for all $1 \leq i \leq n$; and*
- *$(\mathcal{Z}(\boldsymbol{t}), F) \in E^{\mathcal{I}}$ for some $F \in S^{\mathcal{I},\mathcal{Z}}$, for every atom $E(\boldsymbol{t})@S$ occurring in q.*

A BCQ$_@$ q is entailed *by $\mathcal{O}$, written $\mathcal{O} \models q$, iff q is satisfied by every model of $\mathcal{O}$. A BCQ$_@$ that consists of a single atom is an* attributed boolean atomic query *(BAQ$_@$). We say that a BCQ$_@$ is* ground *if it contains only ground specifiers.*

BCQ$_@$ can express conditions on annotations, for instance require that there exists an annotation set where a given attribute is present or has a specific value.

Example 3. *We modify $\mathcal{O}_{\mathsf{ex}}$ to express that those who have a spouse are married, associated with the same annotations:*

$$\exists \mathsf{spouse}@X \sqsubseteq \mathsf{Married}@X.$$

We also add assertions stating that Zsa Zsa Gabor was married to Jack Ryan from 1975 *to* 1976*, while Liz Taylor was married to Richard Burton from* 1975 *to* 1976*, as well as the sources of this information:*

$$\mathsf{spouse}(\mathsf{gabor}, \mathsf{ryan})@[\mathsf{start} : 1975, \mathsf{end} : 1976, \mathsf{src} : \mathsf{s}_1],$$
$$\mathsf{spouse}(\mathsf{gabor}, \mathsf{ryan})@[\mathsf{start} : 1975, \mathsf{end} : 1976, \mathsf{src} : \mathsf{s}_2],$$
$$\mathsf{spouse}(\mathsf{taylor}, \mathsf{burton})@[\mathsf{start} : 1975, \mathsf{end} : 1976, \mathsf{src} : \mathsf{s}_3].$$

The following query expresses that Gabor and Taylor were married (to someone) with the same start and end dates:

$$q_{\mathsf{ex}} = \exists xy \; \mathsf{Married}(\mathsf{gabor})@\lfloor \mathsf{start} : x, \mathsf{end} : y \rfloor \wedge \mathsf{Married}(\mathsf{taylor})@\lfloor \mathsf{start} : x, \mathsf{end} : y \rfloor.$$

By the semantics of DL-Lite$^{\mathcal{R}}_{@}$, it follows that $\mathcal{O}_{\mathsf{ex}} \models q_{\mathsf{ex}}$. This other query expresses that a set of sources that includes s_1 and is associated with Gabor's married status is also associated with Taylor's married status:

$$q'_{\mathsf{ex}} = X : \lfloor \mathsf{src} : \mathsf{s}_1 \rfloor \; \mathsf{Married}(\mathsf{gabor})@X \wedge \mathsf{Married}(\mathsf{taylor})@\lfloor \mathsf{src} : X.\mathsf{src} \rfloor.$$

By the semantics of DL-Lite$^{\mathcal{R}}_{@}$, it follows that $\mathcal{O}_{\mathsf{ex}} \not\models q'_{\mathsf{ex}}$.

While BCQ entailment is NP-complete in DL-Lite$_{\mathcal{R}}$, it follows from Theorem 1 that BAQ$_@$ entailment is already PSPACE-hard. Indeed, satisfiability can be reduced to BAQ$_@$ entailment: $\mathcal{O}$ is unsatisfiable iff $\mathcal{O} \models A(a)$ where A and a are respectively a concept and an individual name that do not occur in $\mathcal{O}$. We show PSPACE-completeness of BCQ$_@$ entailment by describing how to decide $\mathcal{O} \models q$ for a BCQ$_@$ q, using only polynomial space w.r.t. the size of $\mathcal{O}$ and q.

The main ingredients to prove our result are grounding, translation to DL-Lite$_{\mathcal{R}}$, and also *query rewriting*, a prominent query answering technique for DL-Lite$_{\mathcal{R}}$ in which the query is rewritten w.r.t. the concept and role inclusions, to be evaluated over the assertions as in the classical database setting. However, as the ground version of $\mathcal{O}$ is of exponential size and the number of rewritten queries is exponential, we do not compute them but instead guess the DL-Lite$_{\mathcal{R}}$ translation $\mathsf{dl}(q_{\mathcal{Z}})$ of a grounded version $q_{\mathcal{Z}}$ of q together with one of its rewritings q'. We can verify in NP that q' is entailed by the DL-Lite$_{\mathcal{R}}$ translation of the assertions of $\mathcal{O}$, in PTIME that $\mathsf{dl}(q_{\mathcal{Z}})$ is the DL-Lite$_{\mathcal{R}}$ translation of a grounded version of q,

and in PSPACE that q' is indeed a rewriting of $\mathsf{dl}(q_{\mathcal{Z}})$. For this last step, we propose a non-deterministic adaptation of the rewriting algorithm PerfectRef for DL-Lite$_{\mathcal{R}}$ by Calvanese et al. (2007) that takes as input $\mathsf{dl}(q_{\mathcal{Z}})$, q' and $\mathcal{O}$. The main idea is to rewrite $\mathsf{dl}(q_{\mathcal{Z}})$ by guessing at each step an atom of the query together with a positive inclusion that would appear in the DL-Lite$_{\mathcal{R}}$ translation of the grounding of $\mathcal{O}$, thus avoiding the computation of the grounding of $\mathcal{O}$.

Theorem 3. *In DL-Lite$^{\mathcal{R}}_{@}$, BCQ$_{@}$ entailment is in* PSPACE.

The result of Theorem 3, which is for combined complexity, contrasts with the EXPTIME-hardness w.r.t. *data complexity* (only w.r.t. the data size) for MARPL, an attributed logic based on Datalog (Marx, Krötzsch, and Thost 2017). Finally, we show lower complexity bounds in the case of *ground* ontologies. Indeed, when $\mathcal{O}$ is ground, one can build a DL-Lite$_{\mathcal{R}}$ ontology of polynomial size w.r.t. the size of $\mathcal{O}$ that entails the DL-Lite$_{\mathcal{R}}$ translation of a grounded version of q if and only if $\mathcal{O} \models q$.

Theorem 4. *For ground DL-Lite$^{\mathcal{R}}_{@}$ ontologies, satisfiability is in* PTIME *and BCQ$_{@}$ entailment is* NP*-complete.*

Querying Using Provenance Semirings

In this section, we investigate attributed DL in light of provenance semirings (Green, Karvounarakis, and Tannen 2007) and enhance the semantics of DL-Lite$^{\mathcal{R}}_{@}$ to deal with provenance information. Semirings generalize formalisms such as why-provenance, lineages used in view maintenance, or the lineage used by the Trio uncertain management system (Senellart 2017). The main motivation is to use annotations to answer questions such as "Where does the result come from?". Assuming that facts are annotated with their sources, we want to know which combinations of sources lead to the entailment of a query. Such annotations may represent various types of information, such as trust, probability, multiplicity or data classification (see Example 8).

Example 4 (Example 3 cont'd)**.** *The result of the query q_{ex} over the ontology $\mathcal{O}_{\mathsf{ex}}$ can be obtained from source s_3 together with any of $\mathsf{s}_1, \mathsf{s}_2$. Provenance semirings can formalize this information in the form of a* provenance polynomial*:*

$$(\mathsf{s}_1 + \mathsf{s}_2) \times \mathsf{s}_3.$$

The intuitive meaning is that $+$ corresponds to *alternative* use of data and $\times$ to *joint* use of data. The goal of this section is to embed the formalism of provenance semirings into the semantics of DL-Lite$^{\mathcal{R}}_{@}$, so that we can associate annotations using provenance polynomials to queries (e.g., associate the annotation $\mathsf{src} : (\mathsf{s}_1 + \mathsf{s}_2) \times \mathsf{s}_3$ to the query q_{ex} of Example 3).

We define DL-Lite$^{\mathcal{R}}_{@,\mathbb{K}}$ as an order-sorted version of DL-Lite$^{\mathcal{R}}_{@}$. Elements of different sorts correspond to sets of *individual names* $\mathsf{N_I}$, *provenance sums* $\mathsf{N_S}$ and *provenance polynomials* $\mathsf{N_P}$. We represent provenance polynomials with the *positive algebra provenance semiring* for $\mathsf{N_I}$, defined as the commutative semiring of polynomials with variables in $\mathsf{N_I}$ and coefficients from $\mathbb{N}$, with operations defined as usual: $\mathcal{K} = (\mathbb{N}[\mathsf{N_I}], +, \times, 0, 1)$. We denote by $\mathsf{N_P}$ the set of polynomials of $\mathcal{K}$ and by $\mathsf{N_S}$ the subset of $\mathsf{N_P}$ containing the sums of the commutative monoid $(\mathbb{N}[\mathsf{N_I}], +, 0)$. We thus have $\mathsf{N_I} \subseteq \mathsf{N_S} \subseteq \mathsf{N_P}$. We may use the symbols $\sum$ and $\prod$ to denote sum and product of elements in $\mathsf{N_P}$ (which will then also be in $\mathsf{N_P}$). Elements of $\mathsf{N_S}$ are used as values in the ontology specifiers while elements of $\mathsf{N_P}$ appear as values in the query specifiers. Non-linear polynomials indicate the use of several assertions to derive a query, while provenance sums indicate that a query can be derived from different sources.

Role and concept inclusions in DL-Lite$^{\mathcal{R}}_{@,\mathbb{K}}$ are defined similarly as in DL-Lite$^{\mathcal{R}}_{@}$, with the only difference that we allow elements from $\mathsf{N_S}$ to be values of attributes in the specifiers. Concept and role assertions are defined as in DL-Lite$^{\mathcal{R}}_{@}$. The fact that we do not allow values from $\mathsf{N_S}$ in the assertions does not change the expressivity of DL-Lite$^{\mathcal{R}}_{@,\mathbb{K}}$, since inclusions can enforce the entailments of such assertions.

Example 5. *The following concept inclusion restricts that of Example 3 by further requiring that the fact that someone has a spouse has to be associated both with s_1 and with s_2 to conclude that this person is married.*

$$X : \lfloor \mathsf{src} : \mathsf{s}_1 + \mathsf{s}_2 \rfloor \quad (\exists \mathsf{spouse}@X \sqsubseteq \mathsf{Married}@X)$$

Provenance Semantics. We now introduce the semantics of DL-Lite$^{\mathcal{R}}_{@,\mathbb{K}}$, based on provenance sums. A *provenance-interpretation* $\mathcal{I} = (\Delta^{\mathcal{I}}, \cdot^{\mathcal{I}})$ is such that $\cdot^{\mathcal{I}}$ maps polynomials a and b in $\mathsf{N_P}$ to the same element $a^{\mathcal{I}} = b^{\mathcal{I}}$ in $\Delta^{\mathcal{I}}$ if and only if they are mathematically equal[2]. We denote by $\Delta^{\mathcal{I}}_I$ the domain of individuals and by $\Delta^{\mathcal{I}}_S$ the domain of provenance sums, which are the subsets of $\Delta^{\mathcal{I}}$ corresponding to the image of elements in $\mathsf{N_I}$ and $\mathsf{N_S}$, respectively. Thus $\Delta^{\mathcal{I}}_I \subseteq \Delta^{\mathcal{I}}_S \subseteq \Delta^{\mathcal{I}}$. To capture the semantics of provenance sums we develop a notion of closure. Intuitively, if a fact is annotated with n sources then it should also be annotated with the sum of *any subset* of these sources, since the fact can be retrieved alternatively by any source from this subset. For example, assume $(\boldsymbol{a}, F_1), \ldots, (\boldsymbol{a}, F_n)$ are in the interpretation of a concept or a role name E. If there is $(\mathsf{src}^{\mathcal{I}}, \mathsf{s}_i^{\mathcal{I}})$ in each F_i and these annotation sets only differ by such pairs, then for each subset of $\{\mathsf{s}_1, \ldots, \mathsf{s}_n\}$, the interpretation of E should have $(\boldsymbol{a}, F_{\mathsf{s}})$ with F_{s} differing from F_i only by the pair $(\mathsf{src}^{\mathcal{I}}, \mathsf{s}^{\mathcal{I}})$, where s is the sum of the elements of the subset.

We say that $G, H \in \Phi^{\mathcal{I}}$ are *differentiated by* p in F if

$$F = G \setminus \{(p, a) \mid (p, a) \in G\} = H \setminus \{(p, b) \mid (p, b) \in H\}.$$

In this case, we denote by $G +^p H$ the set

$$F \cup \{(p, (a+b)^{\mathcal{I}}) \mid \{a, b\} \subseteq \mathsf{N_P}, (p, a^{\mathcal{I}}) \in G, (p, b^{\mathcal{I}}) \in H\}.$$

A sum of possibly more than two annotation sets differentiated by p may be denoted by $\sum^p_{1 \leq i \leq n} G_i$ and is unique by the commutative law. For $E \in \mathsf{N_C} \cup \mathsf{N_R}$, and $\boldsymbol{a}$ a tuple of the arity of E, we say that $G +^p H$ is *non-primitive* for $\boldsymbol{a}$ and $E^{\mathcal{I}}$ if $\{(\boldsymbol{a}, G), (\boldsymbol{a}, H)\} \subseteq E^{\mathcal{I}}$. We denote by $E^p_{\mathcal{I}, \boldsymbol{a}, F}$ the set of annotation sets G pairwise differentiated by p in $F \in \Phi^{\mathcal{I}}$ such that $(\boldsymbol{a}, G) \in E^{\mathcal{I}}$ with G primitive for $\boldsymbol{a}$ and $E^{\mathcal{I}}$.

Definition 2 (Closure of $E^{\mathcal{I}}$)**.** *$E^{\mathcal{I}}$ is closed under sum if for all tuples $\boldsymbol{a}$ (in $\Delta^{\mathcal{I}}$ or $\Delta^{\mathcal{I}} \times \Delta^{\mathcal{I}}$ according to the arity of E),*

[2] According to associative, commutative and distributive laws. E.g., $(a+b)^{\mathcal{I}} = (b+a)^{\mathcal{I}}$ by the commutative law.

$\{(\boldsymbol{a}, \sum^p_{G\in\sigma} G) \mid \sigma \subseteq E^p_{\mathcal{I},\boldsymbol{a},F}, \sigma \neq \emptyset\} \subseteq E^{\mathcal{I}}$ *for every* $p \in \Delta^{\mathcal{I}}$ *and every* $F \in \Phi^{\mathcal{I}}$.

A provenance-interpretation $\mathcal{I} = (\Delta^{\mathcal{I}}, \cdot^{\mathcal{I}})$ is *well-founded* if $E^{\mathcal{I}}$ is closed under sum for all $E \in \mathsf{N_C} \cup \mathsf{N_R}$. For all $E \in \mathsf{N_C} \cup \mathsf{N_R}$ and $\boldsymbol{a}$ with elements in $\Delta^{\mathcal{I}}$, we also require that the *support* of $E^{\mathcal{I}}$ and $\boldsymbol{a}$ defined as $\{F \mid (\boldsymbol{a}, F) \in E^{\mathcal{I}}\}$ is finite. This ensures that the sum in Definition 2 is finite. An interpretation of DL-Lite$^{\mathcal{R}}_{\mathbb{Q},\mathbb{K}}$ is a well-founded provenance-interpretation. We denote by $\mathbf{S_S}$ the set of specifiers defined in the same way as $\mathbf{S}$ except that we use $\mathsf{N_S}$ instead of $\mathsf{N_I}$ when defining values of attributes. The semantics of specifiers in $\mathbf{S_S}$ is defined as expected following the definition given in the Section 'Attributed DL-Lite' and we use the same notions of satisfiability and entailment. In Definition 2 we consider all subsets of $E^p_{\mathcal{I},\boldsymbol{a},F}$ rather than the sum of its elements. This is to ensure monotonicity of DL-Lite$^{\mathcal{R}}_{\mathbb{Q},\mathbb{K}}$. Otherwise, given for example $A(a)@[p:a]$ and $A(a)@[p:b]$ we would lose the entailment $A(a)@[p:a+b]$ by adding $A(a)@[p:c]$.

Example 6. *Consider the ontology* $\mathcal{O}$ *with the assertions* $\mathsf{spouse}(\mathsf{gabor},\mathsf{ryan})@[\mathsf{src}:\mathsf{s}_1]$, $\mathsf{spouse}(\mathsf{gabor},\mathsf{ryan})@[\mathsf{src}:\mathsf{s}_2]$ *and the concept inclusion of Example 5. Let* $\mathcal{I}$ *have domain* $\Delta^{\mathcal{I}} = \{\mathsf{gabor},\mathsf{ryan},\mathsf{src},\mathsf{s}_1,\mathsf{s}_2,\mathsf{s}_1+\mathsf{s}_2\}$, *interpret each individual name by itself,* $(\mathsf{s}_1+\mathsf{s}_2)^{\mathcal{I}} = \mathsf{s}_1+\mathsf{s}_2$, *and*

$$\mathsf{spouse}^{\mathcal{I}} = \{(\mathsf{gabor},\mathsf{ryan},G),(\mathsf{gabor},\mathsf{ryan},H),(\mathsf{gabor},\mathsf{ryan},G+^{\mathsf{src}}H)\}$$

$$\mathsf{Married}^{\mathcal{I}} = \{(\mathsf{gabor},G+^{\mathsf{src}}H)\}$$ *where* $G = \{(\mathsf{src},\mathsf{s}_1)\}$, $H = \{(\mathsf{src},\mathsf{s}_2)\}$ *and* $G+^{\mathsf{src}}H = \{(\mathsf{src},\mathsf{s}_1+\mathsf{s}_2)\}$.

$\mathsf{spouse}^{\mathcal{I}}$ *and* $\mathsf{Married}^{\mathcal{I}}$ *are closed under sum,* $\mathcal{I}$ *is a model of* $\mathcal{O}$ *and* $\mathcal{O} \models \mathsf{spouse}(\mathsf{gabor},\mathsf{ryan})@\lfloor \mathsf{src}:\mathsf{s}_1+\mathsf{s}_2\rfloor$.

We denote by $\mathbf{S_P}$ the set of specifiers defined in the same way as $\mathbf{S_S}$ except that we use $\mathsf{N_P}$ instead of $\mathsf{N_S}$ for the values of attributes. The semantics of specifiers in $\mathbf{S_P}$ is as expected from the Section 'Attributed DL-Lite'. We assume that all polynomials occurring in a specifier in $\mathbf{S_P}$ are of the form $\Sigma_{1\leq i\leq n_1}\Pi_{1\leq j\leq n_2} a_{i,j}$, where all $a_{i,j} \in \mathsf{N_I}$. Given an interpretation $\mathcal{I} = (\Delta^{\mathcal{I}}, \cdot^{\mathcal{I}})$ and $\{F, G\} \subseteq \Phi^{\mathcal{I}}$, let $F \times G$ be:

$$\{(p,(a\times b)^{\mathcal{I}}) \mid \{a,b\} \subseteq \mathsf{N_P}, (p,a^{\mathcal{I}}) \in F, (p,b^{\mathcal{I}}) \in G\}.$$

Unlike $+^p$, $\times$ is not parameterized by an attribute because products combine different information, whereas sums represent alternative ways of obtaining the same information (i.e., tuple plus the same other attribute-value pairs). A product of annotation sets may be denoted by $\prod_{1\leq i\leq n} G_i$. We next define semiring attributed queries, which allow a ground specifier to be associated to the whole conjunction of atoms.

Definition 3 (Semiring Attributed Queries)**.** *A* semiring attributed boolean conjunctive query *(BCQ*$_{\mathbb{Q},\mathbb{K}}$*) is an expression of the form:*

$$\exists \boldsymbol{x}.X_1 : S_1, \ldots, X_n : S_n\ (\varphi(\boldsymbol{x}))@S,$$

where S *is a ground specifier in* $\mathbf{S_P}$, *for* $1 \leq i \leq n$, $X_i \in \mathsf{N_U}$ *are the set variables occurring in* $\varphi(\boldsymbol{x})$ *and* $S_i \in \mathbf{S_S}$, *and*

$$\varphi(\boldsymbol{x}) = \bigwedge_{1\leq j\leq m} E_j(\boldsymbol{t_j})@T_j$$

where for $1 \leq j \leq m$, $T_j \in \mathbf{S_S}$, $E_j \in \mathsf{N_C} \cup \mathsf{N_R}$ *and* $\boldsymbol{t_j}$ *is a tuple of elements from* $\mathsf{N_I} \cup \boldsymbol{x}$.

If $S = \lfloor\ \rfloor$, *we say that the BCQ*$_{\mathbb{Q},\mathbb{K}}$ *is* plain.

Given a BCQ$_{\mathbb{Q},\mathbb{K}}$ q, *let* q' *be the BCQ*$_{\mathbb{Q}}$ *that results from removing the outer specifier from* q. *Let* $\mathcal{I} = (\Delta^{\mathcal{I}}, \cdot^{\mathcal{I}})$ *be an interpretation and let* $\nu_{\mathcal{I}}(q')$ *be the set of all variable assignments* $\mathcal{Z}$ *that fufill the conditions of Definition 1 for* $\mathcal{I} \models q'$. $\mathcal{I}$ *satisfies* q, *written* $\mathcal{I} \models q$, *if there is a non-empty* $\chi \subseteq \nu_{\mathcal{I}}(q')$ *such that:*

1. *for any* $\mathcal{Z}, \mathcal{Z}' \in \chi$, *there exists* $X \in \mathsf{N_U}$ *occurring in* q *such that* $\mathcal{Z}(X) \neq \mathcal{Z}'(X)$ *or there exists* $x \in \boldsymbol{x}$ *such that* $\mathcal{Z}(x) \neq \mathcal{Z}'(x)$;
2. *for each* $\mathcal{Z} \in \chi$ *and* $1 \leq j \leq m$, *we have that* $(\mathcal{Z}(\boldsymbol{t_j}), F^{\mathcal{Z}}_j) \in E^{\mathcal{I},\mathcal{Z}}_j$ *for some* $F^{\mathcal{Z}}_j \in T^{\mathcal{I},\mathcal{Z}}_j$;
3. *there is* $p \in \Delta^{\mathcal{I}}$ *and* $G \in \Phi^{\mathcal{I}}$ *such that all* $H^{\mathcal{Z}} = \prod_{1\leq j\leq m} F^{\mathcal{Z}}_j$ *with* $\mathcal{Z} \in \chi$ *are differentiated by* p *in* G, *and* $\sum^p_{\mathcal{Z}\in\chi} H^{\mathcal{Z}} \in S^{\mathcal{I},\mathcal{Z}}$.

Essentially, Definition 3 says that: (1) there are different variable assignments which (2) satisfy the homomorphic conditions and (3) correspond to the interpretation of the outer specifier. Our semiring attributed queries can be easily extended so that the outer specifier has fresh and free object variables. In this case the answer to the query would be the set of provenance polynomials related with the respective attribute and the query. Semiring attributed queries can be used to query a DL-Lite$^{\mathcal{R}}_{\mathbb{Q},\mathbb{K}}$ ontology using provenance polynomials, as we illustrate with the following example.

Example 7 (Example 3 cont'd)**.** *We now modify* q_{ex}, *so that we impose provenance constraints on the result:*

$$\exists xy\,(\mathsf{Married}(\mathsf{gabor})@\lfloor \mathsf{start}:x, \mathsf{end}:y\rfloor \wedge \mathsf{Married}(\mathsf{taylor})@\lfloor \mathsf{start}:x, \mathsf{end}:y\rfloor)@\lfloor \mathsf{src}:\gamma\rfloor$$

where γ *is the polynomial* $(\mathsf{s}_1\times\mathsf{s}_3)+(\mathsf{s}_2\times\mathsf{s}_3)$

By the semantics of DL-Lite$^{\mathcal{R}}_{\mathbb{Q},\mathbb{K}}$, *it follows that* $\mathcal{O}_{\mathsf{ex}} \models q_{\mathsf{ex}}$.

All shared attributes are taken into account when combining the annotations, while the non-shared attributes are irrelevant and lost in the product.

Example 8. *The query* $(\mathsf{Married}(a) \wedge \mathsf{Married}(b))@S$ *with*

$$S = \lfloor \mathsf{src}:\mathsf{s}_1\times\mathsf{s}_2, \mathsf{classif}:\mathsf{public}\times\mathsf{confid}, \mathsf{mult}:2\times 3\rfloor$$

is entailed by $\{\mathsf{Married}(a)@[\mathsf{src}:\mathsf{s}_1, \mathsf{classif}:\mathsf{public}, \mathsf{mult}:2], \mathsf{spouse}(b,c)@[\mathsf{src}:\mathsf{s}_2, \mathsf{classif}:\mathsf{confid}, \mathsf{mult}:3, \mathsf{time}:t]\}$ *and the inclusion of Example 3.*

The fact that a *and* b *are both married is obtained by combining sources* s_1 *and* s_2, *and by having access to both public and confidential information. Note that using inclusions to propagate annotations allows the query derived from assertions with multiplicities 2 and 3 to have multiplicity* 2×3, *as it would be under the bag semantics (Nikolaou et al. 2017).*

When interpreted over provenance-interpretations, ontologies in the DL-Lite$^{\mathcal{R}}_{\mathbb{Q}}$ fragment of DL-Lite$^{\mathcal{R}}_{\mathbb{Q},\mathbb{K}}$ (i.e., without sums) can entail queries with sums, as in Example 9.

Example 9. *Let* $\mathcal{O}$ *be the DL-Lite*$^{\mathcal{R}}_{\mathbb{Q}}$ *ontology*

$$\{A(a)@[p:a], A(a)@[p:b], A@X \sqsubseteq \exists R@X\}.$$

Then the query $\exists xy(R(x,y)@[p:a+b])@\lfloor\rfloor$ *follows from* $\mathcal{O}$ *only under the semirings-based semantics.*

Reasoning in DL-Lite$^{\mathcal{R}}_{@,\mathbb{K}}$

Unfortunately, Theorem 5 shows that provenance sums increase the complexity of the satisfiability problem. The proof is by reduction from the word problem for a polynomially space bounded Alternating Turing Machine (ATM) which is EXPTIME-hard (Chandra, Kozen, and Stockmeyer 1981).

Theorem 5. *In DL-Lite$^{\mathcal{R}}_{@,\mathbb{K}}$, satisfiability is* EXPTIME*-hard.*

The hardness result of Theorem 5 holds even for DL-Lite$^{\mathcal{R}}_{@,\mathbb{K}}$ ontologies without expressions of the form $\exists P$, where P is a role expression. Motivated by this negative result, we investigate restricted cases for query answering. We first show that for the class of DL-Lite$^{\mathcal{R}}_{@,\mathbb{K}}$ ontologies which do not contain inclusions with expressions of the form $\exists P$ on the right side, we can check the entailment of BCQ$_{@,\mathbb{K}}$ via a transformation to ground and plain BCQ$_{@,\mathbb{K}}$s. Given such a DL-Lite$^{\mathcal{R}}_{@,\mathbb{K}}$ ontology $\mathcal{O}$, one can translate a BCQ$_{@,\mathbb{K}}$ q into a set of ground and plain BCQ$_{@,\mathbb{K}}$s $\mathsf{gr\_plain}(\mathcal{O}, q)$ such that $\mathcal{O} \models q$ iff there is some $q_{\mathsf{gp}} \in \mathsf{gr\_plain}(\mathcal{O}, q)$ that is entailed by an equisatisfiable ground ontology.

We can assume w.l.o.g. that if $E_j(\boldsymbol{t_j})@T_j$ occurs in q then $T_j \in \mathsf{N_U}$: if T_j is a specifier one can always replace it by a fresh $X \in \mathsf{N_U}$ and add $X : T_j$ to the prefix of q, that is:

$$q = \exists \boldsymbol{x}.\ X_1 : S_1, \ldots, X_m : S_m\ (\bigwedge_{1 \le j \le m} E_j(\boldsymbol{t_j})@X_j)@S.$$

Assume $\star \in \mathsf{N_I}$ does not occur in $\mathcal{O}$ nor in q and let $\mathsf{N_{P_{min}}}$ be a fixed but arbitrary minimal subset of $\mathsf{N_P}$ such that for each $a \in \mathsf{N_P}$, $\mathsf{N_{P_{min}}}$ contains an element b such that a is mathematically equal to b. Let $\mathcal{I}$ be a DL-Lite$^{\mathcal{R}}_{@,\mathbb{K}}$ interpretation with domain $\Delta^{\mathcal{I}} = \mathsf{N_{P_{min}}}$ and such that $a^{\mathcal{I}} = a$ for every $a \in \mathsf{N_{P_{min}}}$. We say that a variable assignment $\mathcal{Z}$ is *compatible with* q if $\mathcal{Z}(X_j) \in S_j^{\mathcal{I},\mathcal{Z}}, 1 \le j \le m$. Let q' be the result of removing the outer specifier from q. Given a compatible $\mathcal{Z}$, a *$\mathcal{Z}$-image* $\bigwedge_{1 \le j \le m} E_j(\boldsymbol{t_j})@T_j$ of q' is obtained by:

- replacing each X_j with $T_j = [a\!:\!b \mid (a,b) \in \mathcal{Z}(X_j)]$;
- replacing each object variable x by $\mathcal{Z}(x)$;
- if $\star$ occurs in some T_j, replacing $\star$ by $\star_{\mathcal{T}_j}$, where $\mathcal{T}_j$ is the set of attribute-value pairs in T_j that do not contain $\star$.

Given a ground specifier T, let $F_T := \{(a^{\mathcal{I}}, b^{\mathcal{I}}) \mid a : b \text{ occurs in } T\} \in \Phi^{\mathcal{I}}$. We define $\mathsf{gr\_plain}(\mathcal{O}, q)$ as the set of ground plain BCQ$_{@,\mathbb{K}}$s:

$$q_{\mathsf{gp}} = (\bigwedge_{1 \le i \le n} (\bigwedge_{1 \le j \le m} E_j(\boldsymbol{t_j^i})@S_j^i))@\lfloor\ \rfloor$$

where the annotation sets $F_i = \prod_{1 \le j \le m} F_{S_j^i}$ $(1 \le i \le n)$ are such that there exists p such that (i) the F_i are differentiated by p in some annotation set, (ii) each F_i contains some (p, a) with $a \in \mathsf{N_P}$, and (iii) $\sum_{1 \le i \le n}^{p} F_i \in S^{\mathcal{I}}$. Also, $\bigwedge_{1 \le j \le m} E_j(\boldsymbol{t_j^i})@S_j^i$ is a $\mathcal{Z}$-image of q' with attribute-value pairs built from elements of $\mathsf{N_S}$. By construction, q_{gp} does not contain variables.

Example 10 (Example 3 cont'd)**.** *The query below is a ground and plain version of the query in Example 7 which is entailed by $\mathcal{O}_{\mathsf{ex}}$.*

$$\mathsf{Married(gabor)@[start : 1975, end : 1976, src : s_1 + s_2]} \wedge$$
$$\mathsf{Married(taylor)@[start : 1975, end : 1976, src : s_3]}.$$

One can show that, for DL-Lite$^{\mathcal{R}}_{@,\mathbb{K}}$ ontologies $\mathcal{O}$ without expressions of the form $\exists P$ on the right side of inclusions, $\mathcal{O} \models q$ iff there is $q_{\mathsf{gp}} \in \mathsf{gr\_plain}(\mathcal{O}, q)$ such that $\mathcal{O}_{\mathsf{gr}} \models q_{\mathsf{gp}}$, where $\mathcal{O}_{\mathsf{gr}}$ is an equisatisfiable ground ontology, obtained in a way similar to our construction of $\mathsf{gr\_plain}(\mathcal{O}, q)$ but imposing that the image of the variable assignments is over a finite set of individual names defined in terms of $\mathcal{O}$. In the case where $\mathcal{O}$ is ground, we further have a polynomial bound on the size of such q_{gp}.

Lemma 1. *Let q be a BCQ$_{@,\mathbb{K}}$ and let $\mathcal{O}$ be a ground DL-Lite$^{\mathcal{R}}_{@,\mathbb{K}}$ ontology without expressions of the form $\exists P$ on the right side of inclusions. $\mathcal{O} \models q$ iff there is $q_{\mathsf{gp}} \in \mathsf{gr\_plain}(\mathcal{O}, q)$ such that (i) $\mathcal{O}_{\mathsf{gr}} \models q_{\mathsf{gp}}$, (ii) the size of q_{gp} is polynomial in $|q|$ and $|\mathcal{O}|$ and (iii) deciding $q_{\mathsf{gp}} \in \mathsf{gr\_plain}(\mathcal{O}, q)$ is in* PTIME*.*

Lemma 1 does not hold for arbitrary DL-Lite$^{\mathcal{R}}_{@,\mathbb{K}}$ ontologies, as illustrated by Example 11.

Example 11. *Let $\mathcal{O}$ be the DL-Lite$^{\mathcal{R}}_{@,\mathbb{K}}$ ontology $\{A \sqsubseteq \exists R, \exists R^- \sqsubseteq A@[p : b], \exists R^- \sqsubseteq \neg B, B(a), A(a)@[p : b]\}$. Then, $\mathcal{O}$ entails $q = \exists x(A(x)@\lfloor\rfloor)@\lfloor p : b + b\rfloor$, since there would be an R-successor in the anonymous part of the model, but there is no $q_{\mathsf{gp}} \in \mathsf{gr\_plain}(\mathcal{O}, q)$ such that $\mathcal{O} \models q_{\mathsf{gp}}$.*

We now use the polynomial bound in Lemma 1 to show an upper bound for a fragment, called *simple*, where we only allow inclusions of the form $E_1@S \sqsubseteq E_2@T$, with E_1 and E_2 concept/role names and S and T ground specifiers. We establish the complexity of BCQ$_{@,\mathbb{K}}$ entailment from simple ontologies. This case is close to the classical problem of query answering over databases, considered by Green, Karvounarakis, and Tannen (2007). Theorem 6 states that this complexity remains the same as in the database case.

Theorem 6. *BCQ$_{@,\mathbb{K}}$ entailment from a simple DL-Lite$^{\mathcal{R}}_{@,\mathbb{K}}$ ontology is* NP*-complete.*

Proof. Let $\mathcal{O}$ be a simple DL-Lite$^{\mathcal{R}}_{@,\mathbb{K}}$ ontology. We first show that one can decide in NP whether $E(\boldsymbol{a})@S$ is entailed from $\mathcal{O}$, where S is a ground specifier.

Claim 1. Deciding whether $\mathcal{O} \models E(\boldsymbol{a})@S$ is in NP.

Proof of Claim 1 We first guess the set $\mathcal{Q}$ of all atomic queries of the form $E_0(\boldsymbol{a})@T_0$ entailed by $\mathcal{O}$ such that $E_0@T_0$ occurs in $\mathcal{O}$ and an ordering for the entailment of such queries. If T_0 is an open specifier then replace it in $\mathcal{Q}$ by $T_{0,\star}$, defined as the ground closed specifier containing all attribute-value pairs in T_0 plus $\star_{\mathcal{S}} : \star_{\mathcal{S}}$ with $\mathcal{S}$ the set of attribute-value pairs in T_0. We make the usual assumption that individual names of the form $\star_{\mathcal{S}}$ do not occur in $\mathcal{O}$ and $E(\boldsymbol{a})@S$. Denote by $\mathcal{Q}_q$ the subset of $\mathcal{Q}$ containing all atomic queries which preceed q in the ordering. For each guessed query $q = E_0(\boldsymbol{a})@T_0$:

- Denote by F_T the set $\{(a, b) \mid a : b \text{ occurs in } T\}$ for any ground specifier T and let $E_0(\boldsymbol{a})@S_1, \ldots, E_0(\boldsymbol{a})@S_n$ be the assertions and atomic queries in $\mathcal{O} \cup \mathcal{Q}_q$ where E_0 and $\boldsymbol{a}$ occur.

- Guess a tree of annotation sets rooted either in F_{T_0} if T_0 is a closed specifier, or in a superset F of F_{T_0} if T_0 is an open specifier, where each non-leaf node F is the parent of children $G_1, \ldots, G_m$ such that $F = \sum_{1 \leq i \leq m}^{p} G_i$, for some attribute p, and such that each leaf is either: one of $F_{S_1}, \ldots, F_{S_n}$, or some F_T (or $F_{T_\star}$ if T is open) such that there exist $E_1@T_1 \sqsubseteq E_0@T$ and $E_1(\boldsymbol{a})@T_1$ (or $E_1(\boldsymbol{a})@T_{1,\star}$ if T_1 is open) in $\mathcal{O} \cup \mathcal{Q}_q$.

Check in polynomial time whether the trees satisfy the described conditions. The size of $\mathcal{Q}$ (and so the number of trees to guess and the size of the ordering) is bounded by the number of atomic queries $E_0(\boldsymbol{a})@T_0$ that can be built from concept/role expressions and individual names in $\mathcal{O}$, so it is polynomial in the size of $\mathcal{O}$.

To check whether $\mathcal{O} \models E(\boldsymbol{a})@S$, we check whether $E(\boldsymbol{a})@S \in \mathcal{Q}$ (assuming w.l.o.g. that $E@S$ occurs in $\mathcal{O}$). The size of each guessed tree is polynomial in the size of $\mathcal{O}$ since each leaf corresponds to an assertion/atomic query in $\mathcal{O}$ or $\mathcal{Q}$ (or an assertion/atomic query in $\mathcal{O}$ or $\mathcal{Q}$ together with an inclusion in $\mathcal{O}$) and they do not repeat in the tree. Thus, one can decide whether $\mathcal{O} \models E(\boldsymbol{a})@S$ in NP.

By Lemma 1, $\mathcal{O} \models q$ iff there exists $q_{\mathsf{gp}} \in \mathsf{gr\_plain}(\mathcal{O}, q)$ such that $\mathcal{O}_{\mathsf{gr}} \models q_{\mathsf{gp}}$. Moreover the size of q_{gp} is polynomial in the size of q and $\mathcal{O}$ and q_{gp} does not contain variables. We thus get the NP upper bound by guessing q_{gp} as well as certificates that $\mathcal{O}_{\mathsf{gr}} \models E(\boldsymbol{a})@S$ for each $E(\boldsymbol{a})@S$ in q_{gp}, using Claim 1 (indeed, $\mathcal{O}_{\mathsf{gr}}$ is also a simple ontology and is of polynomial size w.r.t. $\mathcal{O}$). The lower bound comes from the complexity of BCQ entailment in relational databases. □

One of the difficulties in showing Theorem 6 for arbitrary DL-Lite$^{\mathcal{R}}_{@,\mathbb{K}}$ ontologies is that one can express that elements in the anonymous part of the model are distinct, as illustrated in Example 11, and then our translation does not hold. In this case, $\mathsf{gr\_plain}(\mathcal{O}, q)$ needs to include queries with inequalities to distinguish anonymous elements, and entailment of BCQs with inequalities over DL-Lite$_{\mathcal{R}}$ ontologies easily leads to undecidability (e.g., see Theorem 13 in (Gutiérrez-Basulto et al. 2015)).

We now show an upper bound for *satisfiability* in DL-Lite$^{\mathcal{R}}_{@,\mathbb{K}}$ by translating the ontology into an equisatisfiable ontology in a DL that we call DL-Lite$^{\mathcal{R},\sqcap}_{\mathsf{Horn}}$, which extends DL-Lite$_{\mathcal{R}}$ with *conjunctions on the left side of concept and role inclusions*. Our translation is double-exponential since in DL-Lite$^{\mathcal{R}}_{@,\mathbb{K}}$ we need to ensure, e.g., that elements in the extension of $E@[\mathsf{src}\colon \mathsf{s}_1]$ and $E@[\mathsf{src}\colon \mathsf{s}_2]$ are also in the extension of $E@[\mathsf{src}\colon \mathsf{s}_1 + \mathsf{s}_2]$.

Theorem 7. *In DL-Lite$^{\mathcal{R}}_{@,\mathbb{K}}$, satisfiability is in* 2ExpTime.

Sketch. We first ground the ontology and then translate it into DL-Lite$^{\mathcal{R},\sqcap}_{\mathsf{Horn}}$. We encode the semantics of provenance sums using a double-exponential number of concept and role inclusions with conjunctions on the left side. Since satisfiability in DL-Lite$^{\mathcal{R},\sqcap}_{\mathsf{Horn}}$ is in PTime (Artale et al. 2015) (Theorem 14), the 2ExpTime upper bound follows. □

We next analyse entailment of *plain* BCQ$_{@,\mathbb{K}}$ w.r.t. DL-Lite$^{\mathcal{R}}_{@,\mathbb{K}}$ ontologies: the outer specifier is of the form $\lfloor\ \rfloor$ but inner specifiers can contain provenance sums (as in Ex. 9). We use the fact that BCQ entailment in DL-Lite$^{\mathcal{R},\sqcap}_{\mathsf{Horn}}$ is in NP (Calì, Gottlob, and Pieris 2012, proof of Theorem 3.3).

Theorem 8. *In DL-Lite$^{\mathcal{R},\sqcap}_{\mathsf{Horn}}$, BCQ entailment is in* NP.

Theorem 9 establishes an upper bound for plain queries.

Theorem 9. *In DL-Lite$^{\mathcal{R}}_{@,\mathbb{K}}$, entailment of plain BCQ$_{@,\mathbb{K}}$ is in* N2ExpTime.

Sketch. The proof uses the translation to DL-Lite$^{\mathcal{R},\sqcap}_{\mathsf{Horn}}$ which leads to a double-exponential blowup of the ontology. Here, since queries are plain the translation is as for BCQ$_{@}$s. The result then follows from Theorem 8. □

Conclusion

We investigated the complexity of satisfiability and query answering in attributed DL-Lite$_{\mathcal{R}}$, for both the semantics introduced by Krötzsch et al. (2017) and a new semantics based on provenance semirings, which allows to embed provenance polynomials into the query. In particular, we show that these problems are PSpace-complete for the classical semantics and that in the case of simple ontologies, even query answering under the semirings-based semantics has the same complexity as query answering in DL-Lite$_{\mathcal{R}}$. However, satisfiability of general DL-Lite$^{\mathcal{R}}_{@,\mathbb{K}}$ ontologies is ExpTime-hard.

Related Work. Our attributed ontology language differs from DL-Lite$_{\mathcal{A}}$ (Calvanese et al. 2006), which allows to associate values to individuals or pairs of individuals, rather than to assertions, through binary or ternary relations called attribute concepts or attribute roles. In particular, while we can use the same attribute name to annotate different assertions about the same individual or pair of individuals, it would be ambiguous in DL-Lite$_{\mathcal{A}}$. For instance, we can express that Liz Taylor was married to Richard Burton from 1964 to 1974 and from 1975 to 1976 with $\mathsf{spouse}(\mathsf{taylor}, \mathsf{burton})@[\mathsf{start} : 1964, \mathsf{end} : 1974]$, $\mathsf{spouse}(\mathsf{taylor}, \mathsf{burton})@[\mathsf{start} : 1975, \mathsf{end} : 1976]$, while in DL-Lite$_{\mathcal{A}}$ we would need reification. The query $\mathsf{spouse}(\mathsf{taylor}, \mathsf{burton})@[\mathsf{start} : x, \mathsf{end} : y]$ that returns the start and end dates of the marriages would be more complex (namely, e.g., $\exists z\, \mathsf{spouse}_1(z, \mathsf{taylor}) \wedge \mathsf{spouse}_2(z, \mathsf{burton}) \wedge \mathsf{start}(z, x) \wedge \mathsf{end}(z, y)$). Another difference is the use in DL-Lite$_{\mathcal{A}}$ of two distinct alphabets and interpretation domains for the individuals and the values, following the distinction made in OWL between objects and values.

Regarding provenance, the topic has been extensively studied for relational databases (Cheney, Chiticariu, and Tan 2009), but has also drawn attention in other settings, e.g., for Datalog (Deutch et al. 2014), Datalog$^{+/-}$ (Lukasiewicz et al. 2014), and Semantic Web data, with numerous works proposing provenance models based on semirings for the evaluation of SPARQL queries over annotated RDF, see e.g., (Theoharis et al. 2011; Zimmermann et al. 2012; Geerts et al. 2016). In particular, Zimmermann et al. consider the possibility of having several annotations with different domains (fuzzy, temporal and provenance) and introduce an annotated version of SPARQL that manipulates explicitly

annotations, while most work on provenance only implicitly propagates provenance annotations.

Future Work. Our next step will be the study of the data complexity and the design of practical algorithms for querying attributed DL-Lite ontologies. In particular, we would like to extend the classical DL-Lite rewriting approach to the attributed setting to avoid grounding the ontology. For instance, if an ontology only contains inclusions of the form $E@X \sqsubseteq F@X$, the rewriting algorithm for DL-Lite$_\mathcal{R}$ could be adapted to rewrite an attributed query where annotations sets are propagated in the rewriting process (e.g., $\mathsf{Married}(\mathsf{gabor})@\lfloor \mathsf{start} : 1975 \rfloor$ in Example 3 could be rewritten into $\exists y\, \mathsf{spouse}(\mathsf{gabor}, y)@\lfloor \mathsf{start} : 1975 \rfloor$).

Acknowledgements. We thank Stefan Borgwardt and an anonymous reviewer for pointing us to useful references. Partially supported by ANR-16-CE23-0007-01 ("DICOS").

References

Artale, A.; Calvanese, D.; Kontchakov, R.; and Zakharyaschev, M. 2009. The DL-Lite family and relations. *J. Artif. Intell. Res.* 36:1–69.

Artale, A.; Kontchakov, R.; Ryzhikov, V.; and Zakharyaschev, M. 2015. Tractable interval temporal propositional and description logics. In *Proceedings of AAAI*.

Baader, F.; Calvanese, D.; McGuinness, D.; Nardi, D.; and Patel-Schneider, P., eds. 2007. *The Description Logic Handbook: Theory, Implementation, and Applications*. Cambridge University Press, second edition.

Bienvenu, M., and Ortiz, M. 2015. Ontology-mediated query answering with data-tractable description logics. In *Reasoning Web, Tutorial Lectures*, 218–307.

Bienvenu, M.; Deutch, D.; and Suchanek, F. M. 2012. Provenance for Web 2.0 data. In *Proceedings of Secure Data Management: 9th VLDB Workshop, SDM 2012*.

Bourgaux, C., and Ozaki, A. 2018. Querying attributed DL-Lite ontologies using provenance semirings. Technical Report 02, Free University of Bozen-Bolzano. Available at `https://www.inf.unibz.it/krdb/KRDB%20files/tech-reports/KRDB18-02.pdf`.

Calì, A.; Gottlob, G.; and Pieris, A. 2012. Towards more expressive ontology languages: The query answering problem. *Artif. Intell.* 193:87–128.

Calvanese, D.; De Giacomo, G.; Lembo, D.; Lenzerini, M.; Poggi, A.; and Rosati, R. 2006. Linking data to ontologies: The description logic DL-Lite$_A$. In *Proceedings of the OWLED*06 Workshop on OWL: Experiences and Directions*.

Calvanese, D.; Giacomo, G. D.; Lembo, D.; Lenzerini, M.; and Rosati, R. 2007. Tractable reasoning and efficient query answering in description logics: The DL-Lite family. *J. of Automated Reasoning* 39(3):385–429.

Chandra, A. K.; Kozen, D. C.; and Stockmeyer, L. J. 1981. Alternation. *J. of the ACM* 28(1):114–133.

Cheney, J.; Chiticariu, L.; and Tan, W. C. 2009. Provenance in databases: Why, how, and where. *Foundations and Trends in Databases* 1(4):379–474.

Deutch, D.; Milo, T.; Roy, S.; and Tannen, V. 2014. Circuits for datalog provenance. In *Proceedings of ICDT*.

Geerts, F.; Unger, T.; Karvounarakis, G.; Fundulaki, I.; and Christophides, V. 2016. Algebraic structures for capturing the provenance of SPARQL queries. *J. ACM* 63(1):7:1–7:63.

Green, T. J.; Karvounarakis, G.; and Tannen, V. 2007. Provenance semirings. In *Proceedings of PODS*.

Gutiérrez-Basulto, V.; Ibáñez García, Y.; Kontchakov, R.; and Kostylev, E. V. 2015. Queries with negation and inequalities over lightweight ontologies. *Web Semant.* 35(P4):184–202.

Hoffart, J.; Suchanek, F. M.; Berberich, K.; and Weikum, G. 2013. YAGO2: A spatially and temporally enhanced knowledge base from wikipedia. *Artif. Intell.* 194:28–61.

Krötzsch, M.; Marx, M.; Ozaki, A.; and Thost, V. 2017. Attributed description logics: Ontologies for knowledge graphs. In *Proceedings of ISWC*.

Krötzsch, M.; Marx, M.; Ozaki, A.; and Thost, V. 2018. Attributed description logics: Reasoning on knowledge graphs. In *Proceedings of IJCAI*.

Lukasiewicz, T.; Martinez, M. V.; Predoiu, L.; and Simari, G. I. 2014. Information integration with provenance on the semantic web via probabilistic datalog+/-. In *URSW 2011-2013, Revised Selected Papers*.

Marx, M.; Krötzsch, M.; and Thost, V. 2017. Logic on MARS: Ontologies for generalised property graphs. In *Proceedings of IJCAI*.

Motik, B.; Cuenca Grau, B.; Horrocks, I.; Wu, Z.; Fokoue, A.; and Lutz, C., eds. 2009. *OWL 2 Web Ontology Language: Profiles*. W3C Recommendation. Available at `http://www.w3.org/TR/owl2-profiles/`.

Nikolaou, C.; Kostylev, E. V.; Konstantinidis, G.; Kaminski, M.; Grau, B. C.; and Horrocks, I. 2017. The bag semantics of ontology-based data access. In *Proceedings of IJCAI*.

Rodriguez, M. A., and Neubauer, P. 2010. Constructions from dots and lines. *Bulletin of the American Society for Information Science and Technology* 36(6):35–41.

Savitch, W. J. 1970. Relationships between nondeterministic and deterministic tape complexities. *Journal of Computer and System Sciences* 4(2):177 – 192.

Senellart, P. 2017. Provenance and probabilities in relational databases. *SIGMOD Record* 46(4):5–15.

Suciu, D.; Olteanu, D.; Ré, C.; and Koch, C. 2011. *Probabilistic Databases*. Synthesis Lectures on Data Management. Morgan & Claypool Publishers.

Theoharis, Y.; Fundulaki, I.; Karvounarakis, G.; and Christophides, V. 2011. On provenance of queries on semantic web data. *IEEE Internet Computing* 15(1):31–39.

Vrandečić, D., and Krötzsch, M. 2014. Wikidata: A free collaborative knowledgebase. *Commun. ACM* 57(10).

Zimmermann, A.; Lopes, N.; Polleres, A.; and Straccia, U. 2012. A general framework for representing, reasoning and querying with annotated semantic web data. *J. Web Sem.* 11:72–95.